NEGZ
Kompetenznetzwerk
Digitale Verwaltung

Impulspapier

Model Cards für die Nutzung von KI in der Verwaltung:

Ein Mindeststandard für Transparenz und Vergleichbarkeit

Anna Lena Fehlhaber · Axel Düker

Inhaltliche Ansprechpartner

Anna Lena Fehlhaber fehlhaberlena@gmail.com

Axel Düker axel.dueker@habbel.de

Über das NEGZ

Das NEGZ · Kompetenznetzwerk Digitale Verwaltung ist Fachnetzwerk und Denkfabrik zur Verwaltungsdigitalisierung.

Wir bündeln die Expertise von Unternehmen, Forschungseinrichtungen, öffentlichen Körperschaften und Verbänden, um die Digitalisierung der deutschen Verwaltung zu unterstützen und voranzutreiben.

Wir veröffentlichen Studien und Impulse, veranstalten Austauschformate, vermitteln Kompetenzen und bringen uns in die Fachdiskussion ein.

Impressum

Erscheinungsjahr 2026

ISSN 2626-6032
DOI 10.30418/2626-6032.2026.1

Herausgeber

NEGZ e.V.
Markgrafendamm 24 · 10245 Berlin

030 7543 89 55
office@negz.org · www.negz.org

Gestaltung: Olena Rudman

Titelbild: KI-generiert mit Firefly

Lektorat: Dr. Katharina van Treeck

Auf einen Blick

Künstliche Intelligenz (KI) hält mit hoher Geschwindigkeit Einzug in Wirtschaft und Gesellschaft – und damit auch zunehmend in die Verwaltung. Der Nutzen kann groß sein: von effizienterem Arbeiten bis hin zum Ausbau digitaler Leistungen. Gleichzeitig ist der Einsatz von KI in der öffentlichen Verwaltung besonders anspruchsvoll. Denn wo Verwaltung handelt, müssen Entscheidungen rechtssicher, nachvollziehbar und fair sein.

Den Behörden fehlen jedoch oft Orientierung und Kriterien, wie sie KI-Systeme vertrauenswürdig und rechtssicher einsetzen können. In der Folge sind die technischen und organisatorischen Voraussetzungen uneinheitlich, sodass KI-Lösungen mit unterschiedlichen Qualitäts- und Sicherheitsmaßstäben eingeführt werden. Das erhöht nicht nur die Risiken – etwa im Hinblick auf den Datenschutz oder ungewollte Verzerrungen –, sondern erschwert auch den Vergleich, die Skalierung und die Übertragung von Lösungen zwischen den Kommunen. Gleichzeitig wächst mit der breiten Verfügbarkeit von KI-Werkzeugen der Druck, schnell zu handeln.

Vor diesem Hintergrund schlägt dieses Impulspapier eine pragmatische und wirkungsvolle Lösung vor: Model Cards für KI als verbindlicher Mindeststandard für den Einsatz in der Verwaltung. Model Cards sind kompakte, standardisierte Dokumentationen eines KI-Modells – vergleichbar mit einem Beipackzettel. Sie bündeln die Informationen, die Verwaltungen benötigen, um KI-Modelle zu bewerten – und schaffen so eine verlässliche Grundlage für Beschaffungen, Freigaben und operative Kontrollen. Richtig eingeführt, erleichtern Model Cards es Verwaltungen, eigene Anforderungen zu formulieren, Angebote zu vergleichen und Risiken frühzeitig zu erkennen. Gleichzeitig werden KI-Lösungen anschlussfähig an nationale sowie internationale Standards, wie beispielsweise ISO-Normen oder den Artificial Intelligence Act der EU. So tragen Model Cards dazu bei, digitale Souveränität und öffentliches Vertrauen auch bei breiter Nutzung von KI zu erhalten.

Damit Model Cards ihre Wirkung entfalten können, empfiehlt das Papier eine klare Rollenverteilung der Beteiligten: Anbieter von KI liefern verständliche und vollständige Angaben – nicht nur zu technischen Eckdaten, sondern auch zu Governance, Recht und Betrieb. Verwaltungen definieren Anforderungen, prüfen die vorgelegten Angaben und nutzen diese für die Auswahl, Einführung und das Monitoring der KI. Bund und Länder schaffen Verbindlichkeit durch Leitlinien und Standards. Als nächste Schritte empfiehlt dieses Papier, Model Cards kurzfristig als Mindeststandard in Pilotvorhaben einzusetzen und sie mittelfristig in Beschaffung, Betrieb und Evaluation behördlicher KI-Systeme verbindlich zu verankern.

Inhaltsverzeichnis

Einleitung

Der Vormarsch der KI stellt die öffentliche Verwaltung vor eine doppelte Herausforderung. Kommunen sollen digitale Angebote für Bürger:innen und Wirtschaft ausbauen und effizienter arbeiten. Gleichzeitig müssen sie dabei zentrale Prinzipien des Verwaltungshandelns wahren – Entscheidungen müssen rechtskonform, nachvollziehbar und fair bleiben.

Mit der breiten Verfügbarkeit großer KI-Sprachmodelle wie GPT kommt ein zusätzlicher Treiber hinzu: Viele Mitarbeitende in der Verwaltung nutzen bereits frei verfügbare Systeme, bevor klare Regeln, Freigaben und Zuständigkeiten etabliert sind, und Fachbereiche testen ad hoc, während die externen Erwartungen steigen.

Das trifft auf eine Realität, die ohnehin komplex ist. IT-Infrastrukturen sind meist historisch gewachsen, Daten liegen in unterschiedlichen analogen und digitalen Formaten vor, Schnittstellen sind oft proprietär (das heißt an einen bestimmten Anbieter gebunden) und Sicherheitsniveaus uneinheitlich. Das erschwert Digitalisierung und Kooperation und erhöht Integrationskosten sowie das Risiko, sich langfristig an einzelne Anbieter zu binden. Bei KI-Technologie kommen besondere Vorgaben zur Erklärbarkeit und Transparenz hinzu, die jedoch unter den vorherrschenden Rahmenbedingungen vielerorts nur unzureichend umgesetzt werden können.

Damit KI-Systeme in der Verwaltung vertrauenswürdig und vergleichbar eingesetzt werden können, braucht es gemeinsame Standards. Andernfalls droht ein Flickenteppich aus Einzellösungen und ungleichen Qualitätsniveaus, der Kooperation, Vergleichbarkeit und gegenseitiges Lernen behindert. Standardisierung schafft hier keine Bürokratie, sondern die Grundlage für Sicherheit, Nachvollziehbarkeit und demokratische Kontrolle.

Dieses Impulspapier schlägt daher – als pragmatische Lösung – den verbindlichen Einsatz von Model Cards vor. Model Cards sind kompakte, standardisierte Dokumentationen von KI-Modellen. Sie enthalten die für die Verwaltung entscheidenden Informationen, beispielsweise zu den verwendeten Daten und zur Sicherheit. Damit entsteht eine gemeinsame Grundlage, auf der Verwaltungen KI-Lösungen beschaffen, freigeben und im Betrieb kontrollieren können. Sie erleichtern den Vergleich zwischen Angeboten, machen Risiken früh sichtbar und unterstützen die Einhaltung europäischer und nationaler Vorgaben, etwa im Kontext des Artifical Intelligence Acts (AI Acts) der EU.

Ziel der Einführung von Model Cards ist es, Kommunen und Behörden in ihrer Rolle als prüfende Instanz zu stärken: nicht als Produzenten komplexer KI-Dokumentation, sondern als Anwender standardisierter Informationen, mit denen sie fundierte Entscheidungen treffen können.

Das Papier geht dabei wie folgt vor: Kapitel 2 klärt die zentralen Begriffe und ordnet Model Cards ein. Kapitel 3 erläutert, warum Model Cards gerade für die Verwaltung eine passende Antwort auf aktuelle Herausforderungen darstellen. Kapitel 4 übersetzt den Ansatz in die Praxis: Welche Eigenschaften müssen Model Cards haben, damit sie im Verwaltungsalltag nützlich sind, und welche Inhalte sollten sie konkret abdecken? Kapitel 5 schließt mit Handlungsempfehlungen, wie Model Cards in Verwaltungen eingeführt werden können.

2 Model Cards kurz erklärt

Model Cards sind kompakte, standardisierte Steckbriefe zu KI-Modellen. Sie bündeln zentrale Informationen zu Zweck, verwendeten Daten, Leistungskennzahlen, Einschränkungen sowie Maßnahmen in Bezug auf Datenschutz, Fairness, Sicherheit und Aufsicht. Durch ihre strukturierte Form ermöglichen Model Cards der Verwaltung, ein KI-Modell seriös einzuordnen. Sie geben einen Überblick über Fragen wie:

- **Wofür ist das Modell gedacht? Was ist sein Zweck und Nutzungskontext?**
- **Mit welchen Daten wurde das Modell trainiert?**
- **Welche Risiken und Schwächen sind bekannt?**
- **Wie werden Fairness, Sicherheit und Rechtskonformität gewährleistet?**

Zur Begriffsklärung:

Ein **KI-Modell** ist eine trainierte, mathematisch-algorithmische Struktur, die aus Daten Muster erkennt und auf dieser Grundlage Vorhersagen oder Entscheidungen trifft. Beispiele für solche Modelle sind neuronale Netze, Entscheidungsbäume oder statistische Verfahren, die etwa in der Bilderkennung, Sprachverarbeitung oder in Empfehlungssystemen eingesetzt werden.

Ein **KI-System** umfasst die Gesamtheit aller benötigten Komponenten rund um ein Modell. Neben dem Modell selbst gehören auch die Datenverarbeitung, die Benutzeroberfläche sowie die Infrastruktur, über die das Modell bereitgestellt wird, dazu.

Ein prominentes Beispiel ist ChatGPT: Dieses KI-System basiert auf einem einzelnen großen Sprachmodell (Large Language Model), das mit Milliarden Textbeispielen trainiert wurde, um Sprache zu verstehen und zu erzeugen.

Eine **Model Card** bezieht sich auf das Modell als zentralen Baustein eines KI-Systems. Ursprünglich von Mitchell et al. (2019) vorgeschlagen, haben sich Model Cards in verschiedenen Bereichen als pragmatisches Instrument etabliert, um KI-Modelle verständlich zu dokumentieren.

3 Warum braucht es Model Cards in der Verwaltung?

3.1 KI-Einsatz muss überprüfbar sein

Der größte Nutzen von Model Cards für die öffentliche Verwaltung besteht darin, dass sie aus einer KI-Anwendung ein überprüfbares Produkt machen und so Dialog und Steuerung erleichtern. Hervorzuheben ist insbesondere, dass sie die Nachvollziehbarkeit eines KI-Modells nicht nur technisch, sondern auch rechtlich und organisatorisch abbilden. Wenn Anbieter transparent offenlegen, was das System kann und wo seine Grenzen liegen, kann die Verwaltung fundiert entscheiden, statt sich auf Marketingversprechen zu verlassen.

Dabei beginnt verantwortungsvoller KI-Einsatz bereits vor der Einführung: Personalräte und Datenschutzbeauftragte müssen nachvollziehen können, wie personenbezogene Daten verarbeitet werden und ob Beschäftigte ausreichend geschützt sind. Das ist im öffentlichen Sektor besonders wichtig, weil hier strenge Erwartungen an Rechenschaft, Transparenz und Dokumentation[1] gelten, um das Vertrauen der Bürger:innen zu gewährleisten und sicherzustellen, dass Entscheidungen fair und ethisch getroffen werden. Anonymisiert können Model Cards zudem über Open-Data-Ansätze veröffentlicht werden, um Transparenz und demokratische Kontrolle zu stärken.

Studien zeigen, dass Bürger:innen digitale Angebote befürworten – aber nur, wenn ihre Funktionsweise transparent ist und Fehler vermieden werden.[2] Im Verwaltungskontext bedeutet das: Jede noch so nützliche KI-Anwendung kann scheitern, wenn sie als „Black Box“ wirkt und nicht verantwortungsvoll gesteuert wird. Besonders kritisch sind Anwendungsfehler und Datenschutzverstöße, weil sie zu Reputationsschäden und Vertrauensverlust in die Handlungsfähigkeit der Behörde und damit in die Leistungsfähigkeit und Robustheit der demokratischen Strukturen führen können. Deshalb müssen KI-Systeme, die etwa die Sachbearbeitung unterstützen, indem sie eingereichte Unterlagen und hinterlegte Fachinformationen aus Gesetzen und Verordnungen auswerten, ihre Empfehlungen prüfbar begründen – mit nachvollziehbarer Herleitung und Quellen.

Model Cards sind dabei ausdrücklich kein Ersatz für weitergehende Prüfungen von KI-Anwendungen, wie Folgenabschätzungen, Zertifizierungen oder risikoorientiertes Testen. Durch ihren strukturierten Überblick über die wesentlichen Informationen des verwendeten KI-Modells liegt ihr Wert vielmehr darin, diese Verfahren anschlussfähig und niedrigschwellig zu machen – auch für nichttechnische Entscheidungstragende.

1 Müller et al. (2025)
2 Initiative D21 e.V. und Technische Universität München (2024)

3.2 KI-Einsatz braucht Standards

Zudem tragen Model Cards dazu bei, strukturelle Risiken beim Einsatz von KI zu lindern:

Die KI-Landschaft entwickelt sich schnell, ist aber unübersichtlich. Unterschiedliche Datenformate, proprietäre Schnittstellen und verschiedene Sicherheitsstandards machen es schwer, Systeme miteinander zu verbinden. Das erschwert datengestützte Zusammenarbeit zwischen Institutionen oder Regionen – etwa bei Smart-City-Projekten oder Krisenprävention. Auch bei Dokumentation und Governance von KI gibt es bisher keinen einheitlichen Standard. Anstatt eine einheitliche und vertrauenswürdige digitale Infrastruktur aufzubauen, drohen so Einzellösungen mit stark unterschiedlichen Qualitätsniveaus, die Kooperation und gegenseitiges Lernen ausbremsen. Model Cards schaffen hier ein gemeinsames Fundament, auf dem Verwaltung, Anbieter und Aufsicht sich verständigen und effizienter zusammenarbeiten können.

Außerdem führen fehlende Standards auch zu rechtlicher Unsicherheit. Wenn technische und organisatorische Mindestanforderungen nicht klar sind, geraten Projekte schnell in Grauzonen. Verantwortliche wissen dann oft nicht, was in Bezug auf Datenschutz, Haftung oder Ethik konkret erwartet wird. Das verzögert Projekte, hemmt Investitionen und begünstigt Fehlentwicklungen.

Nicht zuletzt besteht das Risiko sogenannter Lock-in-Effekte. Kommunen, die in proprietäre KI-Lösungen investieren, sind häufig an bestimmte Anbieter gebunden, ohne realistische Möglichkeit eines Wechsels. Fehlende Standards verhindern eine einfache Migration von Daten, Modellen oder Schnittstellen, was langfristig zu technologischer Abhängigkeit und eingeschränkter Innovationsfreiheit führt.

In Medizin und Wirtschaft[3] zeigen etablierte Dokumentationsstandards[4], dass Standardisierung praktikabel ist und Wirkung entfaltet. Auch in Europa gibt es praxistaugliche Vorbilder: Die Niederlande arbeiten etwa mit einem Algorithmenregister[5] – unter vergleichbaren rechtlichen Rahmenbedingungen. Je stärker KI in der öffentlichen Verwaltung skaliert und je mehr interne Datenbestände genutzt werden, desto dringlicher wird es, solche Standards auch hier verbindlich zu verankern.

Genau hier setzen die Model Cards an: Sie lösen nicht alle Probleme – aber sie sind ein wirksamer Schutzmechanismus gegen genau diese unkoordinierte technologische Entwicklung. Sie stellen Vergleichbarkeit her, schaffen pragmatische Standards und machen so den Weg frei für eine verantwortungsvolle und zugleich praktikable Nutzung von KI in der Verwaltung.

3 Konopac & Waltinger, 2021
4 Liang et al. (2024)
5 Algorithmenregister der Niederlande (2025): Die Niederlande betreiben eine öffentliche Datenbank, die alle produktiv eingesetzten KI-Systeme in Bezug auf Funktion, Auswirkung, Datenverarbeitung, Einspruchsrechte und so weiter in verständlicher Sprache dokumentiert.

4 Model Cards in der Verwaltungspraxis

4.1 Worauf es bei Model Cards in der Verwaltung ankommt

Damit Model Cards in der Praxis wirken, müssen sie Fragen beantworten, die für die Verwaltung tatsächlich relevant sind. Es reicht nicht, technische Parameter oder Trainingsmethoden aufzulisten. Entscheidend ist, ob eine Model Card verständlich beantwortet, in welchem Kontext ein KI-Modell genutzt werden kann, welche Daten zugrunde liegen, welche Risiken bestehen und wie Sicherheit sowie Regelkonformität im Betrieb gewährleistet werden.

Zwei Beispiele zeigen, worauf es ankommt:

1) KI-Einsatz zur Priorisierung von Bauanträgen

Wenn eine Kommune die Nutzung eines KI-Systems zur automatisierten Priorisierung von Bauanträgen prüfen möchte, bleibt ohne Model Card offen, ob die Trainingsdaten tatsächlich Bauakten aus vergleichbaren Verfahren enthalten, wie das Modell mit unvollständigen Daten umgeht und ob es Hinweise auf eine systematische Benachteiligung bestimmter Antragsteller gibt.

Eine Model Card kann hingegen zum Beispiel folgende Aspekte offenlegen:

- **Das Modell dient ausschließlich der Sortierung, nicht der Entscheidung.**
- **Es wurde mit Bauakten aus fünf mittleren Großstädten trainiert.**
- **Bekannte Schwächen bestehen bei Anträgen mit sehr untypischen Bauvorhaben.**
- **Eine Bias-Analyse ergab keine systematische Benachteiligung nach Grundstücksgröße oder Eigentümerprofil.**

2) KI-Einsatz zur Überprüfung von Anträgen

Genauso relevant sind Model Cards im Bereich sozialer Leistungen: Wenn ein KI-System Anträge von Bürger:innen vorsortiert, muss klar sein, ob das System nur bei Vollständigkeitsprüfung und Sortierung unterstützt oder auch Entscheidungsempfehlungen gibt, welche Daten einfließen (nur aus dem Antrag oder auch aus externen Registern) und welche Schutzmechanismen verhindern, dass Fehler oder systematische Verzerrungen automatisiert in die Verwaltungsakte einfließen.

Damit solche Fragen zuverlässig beantwortet werden können, braucht es eine standardisierte Struktur. Eine Model Card sollte deshalb immer die wichtigsten Punkte zu Zweck und Nutzungskontext, Datenbasis, bekannten Einschränkungen, Fairness- und Sicherheitsmaßnahmen sowie zu rechtlichen Rahmenbedingungen abdecken. Ihr Vorteil gegenüber akademischen Texten oder aufwendigen Zertifizierungen liegt darin, dass sie diese Informationen kompakt, verständlich und vergleichbar aufbereitet – so, dass sie in der Verwaltungspraxis wirklich genutzt werden können.

4.2 Inhalte einer Model Card: Der Pflichtkern für die Verwaltung

Die folgende Übersicht zeigt beispielhaft, wie die einzelne Rubriken einer Model Card im Verwaltungskontext ausgefüllt sein können:

1. Allgemeine Informationen

- Name und Version des Modells
- Einsatzbereich, zum Beispiel Antragsprüfung, Verkehrsprognose, Risikobewertung oder Texgenerierung
- Verantwortliche Institution/ betreibende Organisation
- Hersteller/Anbieter des Modells
- Kontaktmöglichkeit für Rückfragen

2. Zweck und Nutzungskontext

- Beschreibung des konkreten Einsatzzwecks: Wofür ist das Modell gedacht – und wofür ausdrücklich nicht?
- Begründung der Nutzung, zum Beispiel Effzienz, Entlastung oder Vorhersagefähigkeit
- Zielgruppe und Nutzende
- Einsatzrahmen: Pilotprojekt, produktiv, assistierend oder automatisiert entscheidend
- Erforderliche Fähigkeiten und Kenntnisse der Anwendenden

3. Modellbeschreibung

- Modelltyp und –architektur, zum Beispiel Entscheidungsbaum, neuronales Netz oder Transformer
- Trainingsdaten: Herkunft, Quellen, Zeitraum und Qualität
- Leistungskennzahlen zur Messung der Qualität, Zuverlässigkeit oder Eigenschaften eines Modells
- Limitierungen und bekannte Schwächen, zum Beispiel Generalisierungsfähigkeit[6], Overfitting[7] oder mangelnde Robustheit[8]

4. Datenethik und Fairness

- Analysen möglicher Verzerrungen, zum Beispiel nach Alter, Geschlecht oder ethnischer Zugehörigkeit
- Maßnahmen zur Sicherstellung von Nicht diskrimierung und Fairness[9], zum Beispiel Demographic Parity[10]
- Bewertung von Fehlerraten zwischen Gruppen (sogenannte Disparitätsanalysen[11])

6 Barbiero et al. (2020)
7 Zhang et al. (2024), Institut für Interenet-Sicherheit - if(is) (2025)
8 European Commission (2019), Bundesamt für Sicherheit in der Informationstechnik (2021)
9 Yeom & Tschantz (2019)
10 Fraenkel (2020)
11 Castelnovo et al. (2022)
12 Bundesministerium der Justiz und für Verbraucherschutz (2023)
13 Li et al. (2020), Kumar et al. (2020)
14 Lundberg (2018)

5. Rechtliche und regulatorische Konformität

- Kompatibilität mit der Datenschutz-Grundverordnung (DSGVO): Rechtsgrundlagen, Einwilligungen, Speicherfristen und Ähnliches
- Konformität mit dem AI Act der EU (zum Beispiel Risikokategorie)
- Informationspflichten und Rechte der Betroffenen
- Gestaltung menschlicher Aufsicht, zum Beispiel nach den Konzepten Human-in-the-Loop[12] oder Human-on-the-Loop[13]

6. Transparenz und Nachvollziehbarkeit

- Erklärbarkeit des Modells (zum Beispiel durch Methoden wie SHAP[14], die die Entscheidungen komplexer Modelle nachvollziehbar machen)
- Visualisierung der Entscheidungslogik
- Auditierbarkeit (Prüfbarkeit): Prüfpfade, Logging, Versionierung und Ähnliches

7. Risiken und Schutzmaßnahmen

- Verhalten der KI unter „feindlichen“ Bedingungen, zum Beispiel durch gezielte Manipulationen[15]: Früherkennung durch Monitoring oder Drift Detection im Betrieb[16]
- Maßnahmen zur Resilienz gegenüber Manipulationen, Datenabgriffen und Missbrauch

8. Wartung und Governance

- Zyklen für geplante Updates und Wartungen inklusive Change-Management und Änderungsdokumentation
- Verantwortlichkeiten für Betrieb und Kontrolle
- Externe, unabhängige Prüfungen
- Notfallmechanismen zur sofortigen Deaktivierung bei kritischen Fehlfunktionen

9. Erfahrungen aus dem Praxiseinsatz

- Evidenzbasierte Einschätzungen zur Wirkung, zum Beispiel durch Nutzerfeedback, spezifischen Kennzahlen oder Wirkungsanalysen
- Dokumentierte Fehlfunktionen und daraus abgeleitete Erkenntnisse
- Geplante Verbesserungen oder Nachjustierungen auf Basis von Evaluationen und Praxisrückmeldungen

12 Bundesministerium der Justiz und für Verbraucherschutz (2023)
13 Li et al. (2020), Kumar et al. (2020)
14 Lundberg (2018)
15 Sarker (2023), Costa et al. (2023)
16 Croft (2024), Ribeiro (2025), Lee & Lee (2023)

5 Handlungsempfehlungen: So werden Model Cards Standard

5.1 Klare Rollenverteilung aller Beteiligten

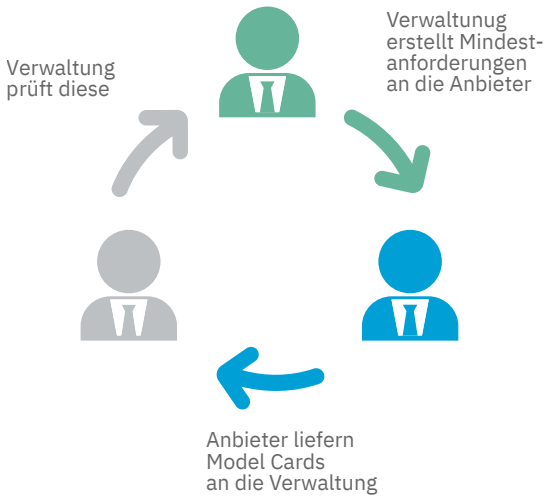
Damit Model Cards wirksam werden, empfiehlt dieses Papier eine klare Rollenverteilung der Agierenden

Anbietende von KI (wie OpenAI, Google, Meta, AWS und Microsoft) stellen vollständige und verständliche Informationen für die Model Cards bereit – nicht nur technische Eckdaten, sondern auch für die Verwaltungen rechtlich und organisatorisch relevante Aspekte.[17] Die Model Cards sollten mindestens technische Merkmale, Datenquellen, rechtliche Konformität, Fairness-Analysen, Sicherheitsmaßnahmen und Governance-Regeln enthalten.

Verwaltungen definieren Mindestanforderungen an die Model Cards, zum Beispiel in Bezug auf Datenschutz, rechtliche Einordnung und Nachvollziehbarkeit. So benötigt eine Bauverwaltung andere Schwerpunkte als ein Sozial- oder Gesundheitsamt, das mit hochsensiblen Personendaten umgeht. Die Behörden bzw. die zuständigen Mitarbeitenden vergleichen und prüfen die vorgelegten Angaben und nutzen sie als Grundlage für Auswahl, Einführung und laufendes Monitoring der KI. Das heißt: Die Verantwortung für vollständige und belastbare Model Cards liegt bei den Anbietern – nicht bei den Behörden. Verwaltungen bringen lediglich ihre Anforderungen ein und können dann prüfen, ob ein Modell diesen Erwartungen entspricht.

Bund und Länder schaffen mit Leitlinien und Referenzprozessen die notwendige Verbindlichkeit. Dabei schließen sie an bestehende Standards und Entwicklungen an: Wichtige Orientierungspunkte sind dabei der AI Act der EU (insbesondere Artikel 13 und Anhang IV), DIN, ISO (insbesondere ISO 42001[18]), IEEE (insbesondere die IEE-7000-Reihe[19]) und nationale Digitalisierungsinitiativen wie FITKO und GovTech Deutschland.

Auf diese Weise entsteht ein pragmatischer Ansatz: Die Hersteller liefern die Inhalte, die Kommunen prüfen diese anhand ihrer eigenen Anforderungen und Bund und Länder schaffen den normativen Rahmen.



17 Dies geschiet in vielen Fällen bereits, allerdings nutzen Verwaltungen sie noch nicht.
18 Die ISO 42001 normiert die Implementierung eines Managementsystems für KI-Systeme, um sicherzustellen, dass sie sicher, ethisch und verantwortungsvoll entwickelt, betrieben und überwacht werden. Die Norm deckt Aspekte wie Risikomanagement, Ethik, Transparenz, Datenmanagement und Sicherheit ab.
19 Die IEEE-7000-Reihe umfasst Standards und Leitfäden zur ethischen Bewertung, Transparenz und Nachvollziehbarkeit von KI-Systemen sowie die Berücksichtigung der Auswirkungen auf die Gesellschaft und die Umwelt.

5.2 Stufenweise Einführung: Von Pilotprojekten zur Verbindlichkeit

Zur Einführung von Model Cards in der Verwaltung bietet sich ein Stufenansatz an:

Kurzfristig

sollten Verwaltungen Model Cards als Mindestanforderung in KI-Pilotvorhabe nutzen, um Erfahrungen zu sammeln. Ergebnisse und Best Practices können in ein nationales oder föderales Transparenregister einfließen, ähnlich dem KI-Transparenzregister des Bundesministerium für Digitalisierung und Staatsmodernisierung[20] oder dem Algorithmenregister der Niederlande.[21]

Mittelfristig

müssen Anbietende verpflichtet werden, vollständige Model Cards bereitzustellen. Dies stellt sicher, dass Model Cards in Beschaffung, Betrieb und Evaluierung behördlicher KI-Systeme genutzt werden können, um fundierte Entscheidungen im Beschaffungsprozess zu treffen.

Langfristig

braucht es einen normativen Rahmen durch Bund und Länder – anschlussfähig an bestehende Standards und europäische Vorgaben. Ein solcher normative Rahmen stellt sicher, dass Model Cards nicht nur in einigen Kommunen als „Best Practice“ existieren, sondern als verbindlicher Standard für alle.

Wenn diese Schritte konsequent umgesetzt werden, entsteht ein klarer Gewinn: Mit einer gemeinsamen Dokumentationsgrundlage lassen sich KI-Systeme nicht nur schneller einführen, sondern auch sicherer und vertrauenswürdiger und leichter zwischen Kommunen übertragen. Model Cards sind dabei kein Allheilmittel, aber ein wichtiger Startpunkt: Sie schaffen Vertrauen, reduzieren Reibungsverluste und stärken die digitale Souveränität der Verwaltung.

20 Bundesministerium für Digitalisierung und Staatsmodernisierung (2025)
21 Algorithm Register Niederlande 2025

Auror:innen

Axel Düker ist seit 2022 als Manager bei der HABBEL GmbH tätig und berät den öffentlichen Sektor in strategischen IT-Themen. Von 2014 bis 2021 war er hauptamtlicher Bürgermeister in Burgwedel und gestaltete die Digitalisierung der Verwaltung. Er bringt umfassende Erfahrung in der Schnittstelle zwischen Verwaltungspraxis, Politik und IT-gestützter Innovation mit.

Anna Lena Fehlhaber ist IT-Sicherheitsforscherin mit Schwerpunkt auf Künstlicher Intelligenz. Sie verfügt über langjährige Erfahrung in der Industrie, insbesondere in der Entwicklung von Foundation-Modellen sowie in Fragen der IT-Sicherheit an der Schnittstelle zu KI. Seit 2024 ist sie im öffentlichen Sektor tätig und verantwortet dort zentrale Themen der KI- und IT-Sicherheit. Parallel engagierte sie sich wissenschaftlich – unter anderem als Dozentin an der Leibniz Universität Hannover und als Gastwissenschaftlerin an diversen weiteren renommierten Forschungseinrichtungen.

Literaturverzeichnis

Algorithm Register Niederlande. 2025. URL: https://algoritmes.overheid.nl/, zuletzt abgerufen am 26.06.2025.

Pietro Barbiero, Giovanni Squillero, Alberto Tonda. 2020. Modeling Generalization in Machine Learning: A Methodological and Computational Study, arxiv:2006.15690 URL: https://arxiv.org/abs/2006.15680, zuletzt abgerufen am 09.07.2025.

Emily Barnes & James Hutson. 2024. Navigating the Complexities of AI: The Critical Role of Interpretability and Explainability in Ensuring Transparency and Trust. Faculty Scholarship 643.

Bundesamt für Sicherheit in der Informationstechnik. 2021. Sicherer, robuster und nachvollziehbarer Einsatz von KI. Probleme, Maßnahmen und Handlungsbedarfe. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf, zuletzt abgerufen am 09.07.2025.

Bundesministerium für Digitalisierung und Staatsmodernisierung. 2025. KI- Transparenzregister. URL: https://maki.beki.bund.de/a/bmi-makimo-app/tabelle, zuletzt abgerufen am 02.09.2025.

Bundesministerium der Justiz und für Verbraucherschutz. 2023. Verwaltungsverfahrensgesetz (VwVfG). § 35 Begriff des Verwaltungsaktes. URL: https://www.gesetze-im-internet.de/vwvfg/__35.html, zuletzt abgerufen am 09.07.2025.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco & Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. Scientific Reports 12, 4209.

Joana C. Costa, Tiago Roxo, Hugo Proença, Pedro R. M. Inácio. 2023. How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses. IEEE Access. 12, S. 1-24.

James Croft. (2024). Identifying drift in ML models: Best practices for generating consistent, reliableresponses.URL: https://techcommunity.microsoft.com/blog/fasttrackforazureblog/identifying-drift-in-ml-models-best-practices-for-generating-consistent-reliable/4040531, zuletzt abgerufen am 09.07.2025.

European Commission. 2024. Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/, zuletzt abgerufen am 09.07.2025.

Initiative D21 e. V. & Technische Universität München. 2024. eGovernment MONITOR. Nutzung und Akzeptanz digitaler Verwaltungsleistungen aus Sicht der Bürger*innen.

[Studienbericht]. URL: https://initiatived21.de/publikationen/egovernment-monitor/2024, zuletzt abgerufen am 09.07.2025.

Institut für Internetsicherheit. 2025. Overfitting im Bereich der Künstlichen Intelligenz (KI). URL: https://vertrauenswuerdigkeit.com/overfitting/, zuletzt abgerufen am 09.07.2025.

Roland Konopac & Ulli Waltinger. 2021. Engineering und IT Tagung 2021, https://www.boeckler.de/pdf/v_2021_09_30_roland_konopac_ulli_waltinger.pdf, Folie 15 ff., zuletzt abgerufen am 26.06.2025.

Sushant Kumar, Sumit Datta, Vishakha Singh, Deepanwita Datta, Sanjay Kumar Singh, and Ritesh Sharma. 2024. Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access 12, S. 75735-75760.

Ye-eun Lee & Tae-jin Lee. 2023. A Study on Efficient AI Model Drift Detection Methods for ML Ops. Journal of Internet Computing and Services 24 (5), S. 17-27.

Nianyu Li, Sridhar Adepu, Eunsuk Kang, and David Garlan. 2020. Explanations for human-on- the-loop: a probabilistic model checking approach. In Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS ,20). Association for Computing Machinery, New York, NY, USA, S. 181–187

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Zric Wu, Yiqun Chen, Daniel Scott Smith & James Zou (2024): Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. Nature Machine Intelligence 6, S. 744–753.

Scott Lundberg. 2018. SHAP documentation. URL: https://shap.readthedocs.io/en/latest/, zuletzt abgerufen am 09.07.2015.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting, FAT Conference 2019. URL: https://arxiv.org/pdf/1810.03993.pdf, zuletzt abgerufen am 26.06.2025.

Oliver Müller, Veronika Lazar, CIO Bund, Leitsatz B6: Transparenz über die behördliche NutzungvonKI. 2025. URL: https://www.cio.bund.de/SharedDocs/faqs/Webs/CIO/DE/digitale-loesungen/KI/3_2_leitsatz6.html, zuletzt abgerufen am 09.07.2025.

Diogo Ribeiro. 2025. Techniques for Monitoring and Managing Model Drift in Production. URL:

https://diogoribeiro7.github.io/machine%20learning/model%20monitoring/techniques_moniitoring_managing_model_drift_production/, zuletzt abgerufen am 09.07.2025.

Iqbal H. Sarker (2023). Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. Security & Privacy 6(5), e295.

Samuel Yeom & Michael Carl Tschantz. 2019. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. URL: https://arxiv.org/pdf/1808.08619v4, zuletzt abgerufen am 09.07.2025.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, Zhumin Chen. 2024. Uncovering Overfitting in Large Language Model Editing. URL: https://arxiv.org/abs/2410.07819, zuletzt abgerufen am 01.10.2025.

www.negz.org